MODELING AND IDENTIFICATION OF ALTERNATIVE FOLDING IN REGULATORY RNAS USING CONTEXT-SENSITIVE HMMS

Byung-Jun Yoon and P. P. Vaidyanathan

Dept. of Electrical Engineering
California Institute of Technology, Pasadena, CA 91125, USA
E-mail: bjyoon@caltech.edu, ppvnath@systems.caltech.edu

ABSTRACT

Recent research on gene regulation has revealed that many noncoding RNAs (ncRNAs) are actively involved in controlling various gene-regulatory networks. For such ncRNAs, their secondary structures play crucial roles in carrying out their functions. Interestingly enough, many regulatory RNAs can choose from two alternative structures based on external factors, which enables the RNAs to regulate the expression of certain genes in an environment-dependent manner. The existence of alternative structures give rise to complex correlations in the primary sequence of the RNA. In this paper, we propose an efficient method for modeling alternative secondary structures in regulatory RNAs. The proposed method can be applied to the prediction of novel regulatory RNAs in genome sequences.

1. INTRODUCTION

Traditionally, proteins have been believed to perform most of the important roles in gene regulation. In the meanwhile, the role of RNAs in gene regulatory networks have remained rather obscure. However, recent research on gene regulation has revealed that many noncoding RNAs (ncRNAs) are actively involved in controlling various genetic networks [1]. These regulatory RNAs include microRNAs (miRNAs) [2], riboregulators [3], riboswitches [4], and many others.

Many functional ncRNAs have well-conserved secondary structures, as these structures are crucial in carrying out their biological functions. Typically, an RNA sequence adopts a single "biologically correct" secondary structure. However, there exist also examples of RNAs that can choose from alternative structures, thereby changing their characteristics. In fact, many regulatory RNAs can make conformational changes depending on one or more environmental cues to regulate the expression level of certain genes [4, 5]. *Riboswitches* are good examples of such RNAs [4]. They are highly structured RNAs that are usually found in the 5' untranslated regions (UTRs) of certain mRNAs. Riboswitches change their secondary structures upon binding to specific metabolites, thereby controlling the expression of the corresponding metabolic genes.

In addition to natural RNAs with differential folding, there are also engineered RNAs that can be used for controlling gene expression based on a similar mechanism. The *antiswitch* designed by Bayer and Smolke is an RNA-based regulator that can directly control the expression of a target transcript in a ligand-dependent manner [6]. Fig. 1 illustrates the general mechanism of an antiswitch regulator. When the effector ligand is absent, the antisense domain in the antiswitch (which is complementary to the target mRNA transcript) is sequestered. As the antiswitch cannot bind to the target, the gene expression of the target is turned on. In the presence of a specific ligand, the antiswitch binds to the ligand, resulting in a change in its secondary structure as shown in Fig. 1 (b). This conformational change releases the antisense domain, thereby allowing the antiswitch to bind to the target mRNA, which is illustrated in Fig. 1 (c). As a result, the antiswitch regulator will suppress the expression of the target gene.

Work supported in parts by the NSF grant CCF-0428326 and the Microsoft Research Graduate Fellowship.

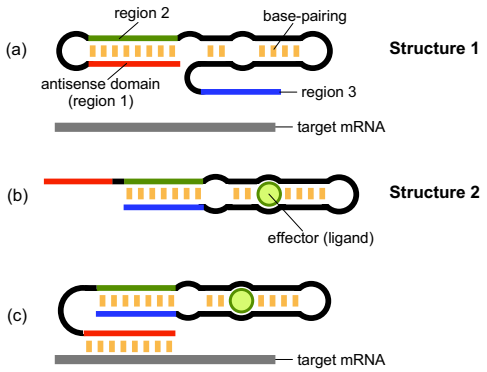

Fig. 1. An antiswitch regulator. (a) Secondary structure of the antiswitch in the absence of ligand. (b) The structure changes upon binding the ligand. (c) In the presence of ligand, the antiswitch can bind to the target mRNA, suppressing its expression.

2. MODELING RNA SEQUENCES WITH ALTERNATIVE SECONDARY STRUCTURES

The existence of alternative secondary structures introduce complex correlations in the primary sequence of the RNA. As an example, let us consider the primary sequence of the antiswitch shown in Fig 2. In the absence of the target ligand, region 1 (the antisense domain) forms base-pairs with region 2. Therefore, there exist correlations between bases in region 1 and those in region 2. When the target ligand is present, region 3 can fold onto region 2, hence there exist also correlations between these two regions. As a result, the bases in region 2 are correlated to the bases in region 1 and also to the bases in region 3. The overall base correlations are depicted in Fig. 2 (a), where the arcs indicate the correlations between bases. Such correlations cannot be modeled using a *stochastic context-free grammar* (SCFG) [7] or a *context-sensitive HMM* (csHMM) [8], and we have to resort to more general grammars such as *context-sensitive grammars* (CSGs). However, CSGs that can represent sequences with correlations shown in Fig. 2 (a) tend to get very complex. Moreover, parsing CSGs is an NP-complete problem, and there is no polynomial-time algorithm that can be used in general [7].

However, we can circumvent these difficulties by adopting the

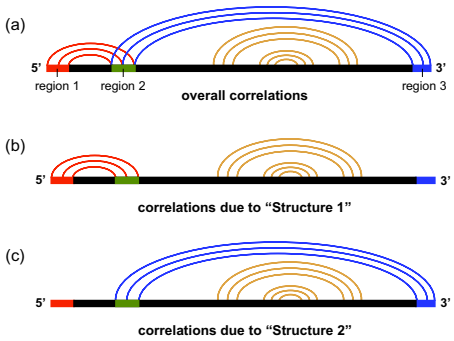

Fig. 2. Base correlations in the primary sequence of an antiswitch. (a) Overall correlations. (b) Correlations due to Structure 1 (in the absence of ligand). (c) Correlations due to Structure 2 (in the presence of ligand).

following strategy. Instead of modeling the overall correlations in the RNA sequence by a single model, we can use multiple csHMMs to represent the respective correlations that arise from each of the alternative RNA secondary structures. For example, we can use a csHMM to describe the correlations that arise from "Structure 1" (shown in Fig. 2 (b)) and use another csHMM to describe the correlations that arise from "Structure 2" (shown in Fig. 2 (c)). Given a novel RNA sequence, we can score it based on the two csHMMs using an efficient polynomial-time algorithm [8]. We can combine the two scores to determine how close the given RNA sequence is to the original RNA sequence modeled by the csHMMs.

3. SIMULATION RESULTS

To demonstrate the efficacy of the proposed method, we constructed two csHMMs as shown in Fig. 3. The csHMM in Fig. 3 (a) models the correlations that arise from "Structure 1", which are shown in Fig. 2 (b). Note that S_n is a single-emission state, P_n is a pairwise-emission state, and C_n is the context-sensitive state that corresponds to P_n. The states P_1 and C_1 work together to generate the antisense-stem that sequesters the antisense domain in the absence of ligand. Similarly, the state-pairs (P_2, C_2) and (P_3, C_3) generate the other stems in "Structure 1". Likewise, the csHMM illustrated in Fig. 3 (b) represents the correlations shown in Fig. 2 (c), which arise from "Structure 2". Based on these csHMMs, we can compute

$$S_1(\mathbf{x}) = \frac{1}{L}\log_2 \frac{P(\mathbf{x}|\text{csHMM1})}{P(\mathbf{x}|H_0)}, S_2(\mathbf{x}) = \frac{1}{L}\log_2 \frac{P(\mathbf{x}|\text{csHMM2})}{P(\mathbf{x}|H_0)}$$

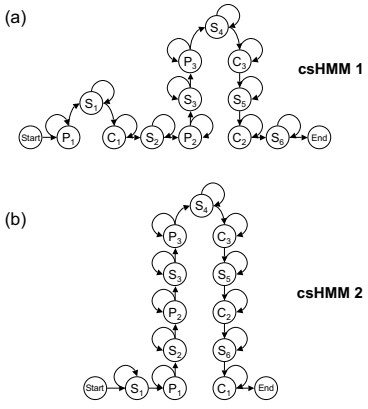

Fig. 3. (a) The csHMM that represents Structure 1. (b) The csHMM that represents Structure 2.

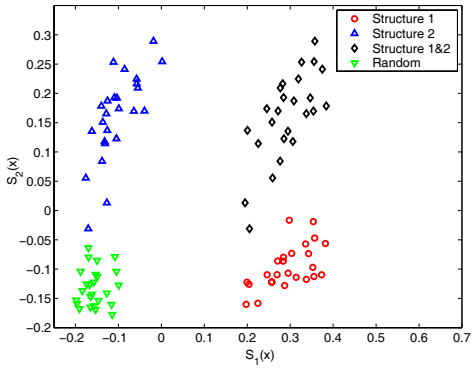

Fig. 4. Plot of $(S_1(\mathbf{x}), S_2(\mathbf{x}))$.

for a given RNA sequence $\mathbf{x}$, where L is the length of $\mathbf{x}$ and H_0 is the random model with i.i.d. assumption. Fig. 4 shows a plot of $(S_1(\mathbf{x}), S_2(\mathbf{x}))$ for 100 test sequences $\mathbf{x}$. As we can see in Fig. 4, sequences with alternative structures (depicted by black diamonds) are well-separated from sequences with either "Structure 1" or "Structure 2", or unstructured sequences.

4. CONCLUDING REMARKS

In this paper, we proposed an efficient method for modeling alternative secondary structures in an RNA molecule. The proposed method uses multiple csHMMs, where each model represents one of the possible structures that can be adopted by the given RNA molecule. This approach effectively discriminates between sequences that can differentially fold and sequences that cannot, at a relatively low (polynomial) computational cost. As many regulatory RNAs can fold to alternative structures, the proposed scheme can be used for finding novel homologues of such RNAs in genome sequences.

5. REFERENCES

[1] S. Gottesman, "Stealth regulation: biological circuits with small RNA switches", *Genes and Development* vol. 16, pp. 2829-2842, 2002.

[2] D. P. Bartel, "MicroRNAs: Genomics, Biogenesis, Mechanism, and Function", *Cell*, vol. 116, pp. 281-297, 2004.

[3] V. A. Erdmann, M. Z. Barciszewska, M. Szymanski, A. Hochberg, N. de Groot, J. Barciszewski, "The non-coding RNAs as riboregulators", *Nucleic Acids Research*, vol. 29, pp. 189-193, 2001.

[4] B. J. Tucker and R. R. Breaker, "Riboswitches as versatile gene control elements", *Current Opinion in Structural Biology*, vol. 15, pp. 342-348, 2005.

[5] T. M. Henkin and C. Yanofsky, "Regulation by transcription attenuation in bacteria: how RNA provides instructions for transcription termination/antitermination decisions.", *Bioessays*, vol. 24, pp. 700-707, 2002.

[6] T. S. Bayer and C. D. Smolke, "Programmable ligand-controlled riboregulators of eykaryotic gene expression", *Nature Biotechnology*, vol. 23, pp. 337-343, 2005.

[7] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis*, Cambridge Univ. Press, Cambridge, UK, 1998.

[8] B.-J. Yoon and P. P. Vaidyanathan, "Scoring algorithm for context-sensitive HMMs with application to RNA secondary structure analysis", *Proc. IEEE Int. Workshop on Genomic Signal Processing and Statistics (GENSIPS)*, Newport, RI, May 2005.